Genome-wide identification of host-segregating SNPs for source attribution of clinical *Campylobacter coli* isolates

Quentin Jehanne[1,3], Ben Pascoe[2], Lucie Bénéjat[1], Astrid Ducournau[1], Alice Buissonnière[1], Evangelos Mourkas[2], Francis Mégraud[1,3], Emilie Bessède[1,3], Samuel K Sheppard[2], Philippe Lehours[1,3,#]

[1] French National Reference Center for Campylobacters & Helicobacters, Bordeaux Hospital University Center, 33300, Bordeaux, France

[2] The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Claverton Down, BA2 7AY, Bath, United Kingdom

[3] Univ. Bordeaux, INSERM, BaRITOn, U1053, F-33000 Bordeaux, France

#Corresponding author: Prof Philippe Lehours, CHU Pellegrin, Laboratoire de Bactériologie, CNR des Campylobacters et des Hélicobacters, Place Amélie Raba Léon, 33076 Bordeaux cedex ; tel: +33557571286 ; mail: philippe.lehours@u-bordeaux.fr

Running title: Source attribution for *Campylobacter coli*

Keywords: *Campylobacter coli,* SNP, source attribution, genomics, genotyping.

ABSTRACT

Campylobacter is among the most common causes of gastroenteritis worldwide. *Campylobacter jejuni* and *Campylobacter coli* are the most common species causing human-disease. DNA-sequence-based methods for strain characterization have focussed largely on *C. jejuni*, responsible for 80-90% of infections, meaning that *C. coli* epidemiology has lagged behind. Here we have analyzed the genome of 450 *C. coli* isolates to determine genetic markers that can discriminate isolates sampled from 3 major reservoir hosts (chickens, cattle and pigs). These markers were then applied to identify the source of infection of 147 *C. coli* from French clinical cases. Using STRUCTURE software, 259 potential host-segregating markers were revealed by probabilistic characterization of SNP frequency variation in strain collections from three different hosts. These SNPs were found in 41 genes or intergenic regions, mostly coding for proteins involved in motility and membrane functions. Source attribution of clinical isolates based on the differential presence of these markers confirmed chicken as the most common source of *C. coli* infection in France.

IMPORTANCE Genome-wide and source attribution studies based on *Campylobacter* species have shown their importance for the understanding of foodborne infections. Although the use of MLST based on 7 genes from *C. jejuni* is a powerful method to structure populations, when applied to *C. coli* results have not clearly demonstrated their robustness. Therefore, we aim here to provide more accurate data based on the identification of single-nucleotide polymorphisms. Results from this study reveal an important number of host-segregating SNPs, found in proteins implied in motility, membrane functions or DNA repair systems. These findings offer new interesting opportunities for further study on *C. coli* adaptation to its environment. Additionally, the results demonstrate that poultry is potentially the main reservoir of *C. coli* in France.

INTRODUCTION

Campylobacter is the leading cause of bacterial gastroenteritis worldwide (1), with around 800,000 campylobacteriosis cases in the USA (2) and 200,000 in the European Union (3) each year. Demographic, dietary and surveillance programs variations have made it difficult to generalise understanding of *Campylobacter* epidemiology to all countries. For example, while there are an estimated 68,000 foodborne infections every year in France (4), the number attributable to *Campylobacter* is not clearly defined, and there are questions about the relative importance of different *Campylobacter* species (5) (6) (7).

C. jejuni and *C. coli* are part of the commensal microbiota of many bird and animal species (8). Human infection typically occurs via consumption of contaminated meat - especially chicken (9) (10) (11), water or direct contact with animals (livestock farming). Infection is usually self-limiting with mild symptoms including abdominal cramps, diarrhoea and fever. However, more severe symptoms such as bloodstream infections and vascular disease can occur, particularly at the extreme ages of life, in immunosuppressed, diabetic or cancer patients, and in rare cases, post-infectious complications include Guillain–Barré syndrome (12) and irritable bowel syndrome (13). Prolonged or severe campylobacteriosis can require the administration of macrolide (azithromycin) or quinolone (ciprofloxacin) (14) (15) antibiotics but increasing resistance, particularly among *C. coli* isolates (16), is reducing treatment options.

C. coli is responsible for an increasing number of infections, accounting for approximately 15% of all campylobacteriosis cases (6). While much research focuses on *C. jejuni,* accounting for about 85% of cases, there are proportional differences between countries potentially reflecting variations in diet

(17) and host source (18) (19). European studies have typically associated *C. coli* with pigs and sheep (5) (20) (21). However, intensive agricultural practices in recent decades have dramatically changed the distribution of livestock species on earth creating opportunities for host transitions (22). This has likely driven changes to the natural host associations of both *C. jejuni* and *C. coli* which are regularly isolated from cattle and chickens (9). This host melting-pot has also dramatically affected the evolution of livestock associated *C. coli* leading to the emergence of a dominant disease-causing *C. coli* lineage, the ST-828 clonal complex (*CC-828*) (23), that has a mosaic genome with over 10% of the genes having been acquired from *C. jejuni* by horizontal gene transfer (24) (25) (26). This genome plasticity is particularly of concern for *C. coli* which acquires antimicrobial resistance genes more easily than *C. jejuni* (14) (16).

Genotyping methods such as multilocus sequence typing (MLST) (27) (28) have improved our understanding of *Campylobacter* population structure, revealing host-specialist and host-generalist lineages (29). This host association has underpinned the development of methods that quantitatively attribute the source of human infections (9) (11). However, rapid host-switching by host generalist *Campylobacter*, including *C. coli CC-828*, can often confound these methods because, for some lineages, strains associated with one host source can be found in another (22) (30). The adoption of whole genome sequencing techniques and availability of curated genome databases (31) have allowed the incorporation of a broader number of host-segregating epidemiological markers in source attribution methods (32) (33). This additional genome information has increased the resolution allowing attribution of invasive/non-invasive strains from poultry (34) as well as geographical attribution of UK/USA isolates (19). However, almost all studies focussed exclusively on *C. jejuni* (35), and no study aimed to specifically identify host-segregating markers in *C. coli* genomes.

89

90 In this study, we analyzed 450 *C. coli* genomes from public databases with defined sampling sources
91 including chickens, cattle and pigs. Using comparative genomics approaches we: (i) tested the ability
92 of traditional MLST-based methods to determine the source of *C. coli* with isolates from known
93 source reservoirs; (ii) identified host-segregating SNPs in *C. coli* genomes; (iii) determined the
94 relative contribution of different *C. coli* infection sources in France. MLST was found to be a good
95 proxy for more complex whole genome SNP-based analysis, showing similar power for segregating
96 isolates from cattle host. However, additional discrimination of isolates from chicken and pig hosts
97 was achieved by identifying genome-wide host-segregating SNPs. In the final probabilistic model,
98 using 259 host-segregating SNPs, chicken was found to be the most common source of *C. coli*
99 infection in France.

100 **RESULTS**

101

102 ***CC-828* isolates segregate by host**

103 From all 3 datasets, Dataset S1, S2 and S3 (cf. material and methods), nearly all isolates belonged to
104 the clonal complex -828 (780 isolates out of 900). The second most common clonal complex
105 identified was *ST-1150* (26) with four isolates, sampled from chicken. From the allelic profiles
106 minimum spanning tree, 3 clusters can be identified corresponding to the source of isolation (Figure
107 1). Cattle isolates clustered together, with 162 isolates (64.8% of all cattle isolates) assigned to *ST-*
108 *1068* (36). Chicken and pig isolates belonged to 78 and 83 sequence types respectively (contrary to
109 cattle with 27 different sequence types), with 24.2% isolates belonging to *STs -828*, *-829*, *-825*, *-854*,
110 and *-1119*. Furthermore, 40.1% of all clinical isolates belonged to *STs -825*, *-827*, *-832* and *-860*.
111 Initial evidence for a role for chicken as a reservoir for human infection was provided by the
112 clustering of clinical isolates together with isolates from chicken on the phylogenetic tree. The
113 second tree constructed using maximum-likelihood approach from concatenated SNPs sequences
114 revealed distinctive partitioning of isolates according to source (Figure 2). *C. coli* isolated from cattle
115 constitute a very distinct cluster; 168 isolates (67.2% of all cattle isolates) are located at the bottom
116 of the tree and belonged to *ST-1068*. Distances were also shorter within cattle population compared
117 to chicken and pig isolates where more variability was observed within both clades. While many
118 clinical isolates clustered among chicken isolates, six clinical isolates were found along a long branch
119 of the chicken's clade - these isolates were interestingly attributed to pig using STRUCTURE (below).
120
121

Host-segregating SNPs differentiate *C. coli* isolated from different hosts

Putative host-segregating SNPs were identified by aligning all 450 isolates selected for marker determination against three *C. coli* reference genomes. Alignment of isolates against the OR12 *C. coli* reference strain identified 283,320 variant sites. In order to remove weakly discriminating polymorphisms, SNP versions represented in more than two thirds of all isolates were filtered, leaving 26,131 variant sites. Similar alignment and filtering performed against the HC2-48 strain resulted in 202,111 variants, filtered to 24,395; and alignment against the ZV1224 reference identified 242,574 SNPs, which were filtered to 20,827. Host-segregating SNPs were identified by performing source attribution tests using each variant individually and all 450 isolates. SNPs with at least 70 % accuracy for at least one source in the self-attribution test included 43, 183 and 33 from each alignment with the OR12, HC2-48 and ZV1224 reference strains, respectively (Table 1). Most of the self-attribution tests showed rates fluctuating between 30% and 40% (Figure 3) (51.2%, 50.5%, 48% of all variants for the chicken, cattle and pig variants, respectively); 33% indicates a complete inability to differentiate 3 individuals. In total, 259 host-segregating SNPs from 41 nucleotide sequences were carried forward for further analyses.

To contextualize host-segregating SNPs within genes, Blast-x annotation identified 32 coding regions for known proteins, 5 hypothetical proteins as well as 4 intergenic regions (Suppl Table S1). Several SNPs (n=27) were found in proteins involved in motility, which plays an important role in bacterial host adaptation: 12 and 4 SNPs in flagellar proteins *FliK* (with 2 SNPs in its basal-body rod modification protein *FlgD*) and *FliD* respectively, known to modulate flagellar hook length (37) and to act as an immunodominant protein (38); 5 SNPs from methyl-accepting chemotaxis proteins (TLP-like protein (39)) or intergenic regions before methyl-accepting chemotaxis proteins; as well as 4

145 SNPs in one aerotaxis receptor belonging to *CetC*, a protein involved in regulating energy taxis (40).
146 Another protein involved in bacterial adaptation to its environment has also been identified from
147 OR12 chicken reference (3 SNPs): *SbmA* (41), a peptide antibiotic transporter described in many
148 gram-negative bacteria. SNPs were also found in proteins involved in metabolism and membrane
149 functions: 3 SNPs from an histidine kinase, 5 SNPs from a single-domain globin protein, known to
150 play a role against NO and nitrosative stress (42), and a *LamB/YcsF* family protein with 5 SNPs. Two
151 phosphate-binding proteins showed the presence of one SNP from OR12 chicken reference variant
152 calling as well as one SNP from ZV1224 pig reference. Proteins involved in DNA activities have also
153 been identified, with a total of 56 SNPs: DNA recombination/repair protein *RecA*, excinuclease ABC
154 subunit C (*UvrC*) (43), two restriction endonucleases from HC2-48 and ZV1224 references, and one
155 transcriptional regulator. Two hypothetical proteins from OR12 and ZV1224 with 11 and 8 host-
156 segregating SNPs respectively have been found to be the same protein: its domains and amino-acid
157 sequence depending the source should be further investigated. Finally, a total of 110 SNPs were
158 within 2 hypothetical proteins (from the HC2-48 cattle reference), which reflected highly variable
159 and isolate-specific regions, and should not be taken into account.
160
161 **Genome-wide host-segregating SNPs provide more accurate source attribution than MLST alleles**
162 The degree of SNP segregation among isolates from different hosts, and hence the potential as
163 marker for source attribution using STRUCTURE, was quantified. Self-attributions of chicken and pig
164 isolates within the marker-determination dataset were consistently correct (Table 2). Using 43 SNPs
165 detected from OR12 alignment as host-segregating markers allowed an average correct self-
166 attribution of 88.35% (s.d. ± 6.2%), 63.75% (s.d. ± 9.2%) and 96.2% (s.d. ± 4.1%) for chickens, cattle
167 and pigs respectively. Using 183 SNPs from HC2-48 alignment correct self-attribution was achieved

for chicken, cattle and pig isolates with 91.05% (s.d. ± 5.7%), 75% (s.d. ± 9.7%) and 42.45% (s.d. ± 18.7%) accuracy respectively, and 74.95% (s.d. ± 13.9%), 19.65% (s.d. ± 10.1%) and 94.65% (s.d. ± 5.2%) for the 33 SNPs from ZV1224 alignment. Low self-attribution rate of cattle isolates using SNPs from pig reference was observed: these isolates were not correctly attributed and were considered as 50% chicken and 50% pig. When using all the SNPs simultaneously (n=259), correct self-attribution showed average scores of 91.95% (s.d. ± 5.86%), 77% (s.d. ± 8.65%) and 95.25% (s.d. ± 4.4%) for chickens, cattle and pigs respectively. This is a considerable improvement of self-attribution using the 7 MLST genes which returned average scores of 73.6% (s.d. ± 9.1%), 76.8% (s.d. ± 9.4%) and 74.4% (s.d. ± 9.5%) for chickens, cattle and pigs respectively. Source attribution of cattle *C. coli* isolates of marker-determination dataset was similar between the two types of markers (genotype or allele) whereas SNPs performed significantly better for chicken and pig populations than the 7 MLST genes. Finally, the discriminatory power of host-segregating SNPs and MLST genes was evaluated performing source re-attribution of 299 *C. coli* isolates from the validation dataset. SNPs showed correct re-attribution proportions of 96.2% (s.d. ± 1.03%), 84% (s.d. ± 0%) and 89% (s.d. ± 0%), and MLST genes scores of 87% (s.d. ± 0%), 81% (s.d. ± 0%) and 65% (s.d. ± 0%) for chicken, cattle and pig populations, respectively (Figure 4). Overall, SNPs were able to better re-attribute *C. coli* marker-determination and validation isolates to their source than MLST genes, more specifically for chickens and pig populations.

Chickens are a major source of *C. coli* infection in France

Source attribution of clinical isolates was performed using MLST alleles and all host-segregating SNPs with correct self-attribution >70% (n=259) in the marker-determination and training dataset using STRUCTURE (Figure 5). Using MLST genes, 89 clinical isolates (60.5%) were attributed to chickens, 13

191 to cattle (9%), 6 to pigs (4%) and 39 clinical isolates (26.5%) showed attribution scores lower than
192 70% and were therefore considered as ”inconclusive attributions”. Inconclusive attributions
193 specifically concern 3 commonly found sequence types: *STs -827*, *-1055* and *-1595*, representing
194 48.7% of inconclusive attributions (n=19). In contrast, using the 259 SNPs, 138 isolates (94%) were
195 attributed to chickens, 9 to pigs (6%) (with an average of source probability equal to 100%) and none
196 to the cattle population. Therefore whatever the approach (MLST or SNPs), a large proportion of *C.*
197 *coli* clinical isolates were attributed to chickens. However, the attribution scores were more variable
198 with MLST (on average around 80%) whereas for the genome-wide host-segregating SNPs, the
199 clinical isolates were more efficiently attributed to their infection source (Table 3).

200 **DISCUSSION**

201

202 The increasing availability of bacterial isolate genome collections and bioinformatics tools for large-
203 scale analysis provides significant opportunities for understanding the genetic basis of phenotype
204 variation in bacteria. Host adaptation is a key feature in the epidemiology of zoonotic pathogens
205 (44), such as *Campylobacter,* and there has been considerable effort to identify host-associated
206 genetic variation that can improve understanding of the evolution and origin of infecting strains.
207 Comparative genomic analyses have revealed core and accessory genome variation within *C. jejuni*
208 that is associated with a given host/environment (45) (46) and this has been used to identify
209 genome-wide host-segregating markers for source attribution (32). However, little comparable work
210 has focussed on *C. coli.*

211

212 Genetic variation in bacterial genomes not only reflects adaptation to different hosts/sources but
213 also temporal and geographic variation among sample collections (19). Some studies avoid the
214 potential confounding effect of phylogeographic variation by using national isolate collections: for
215 example, *Campylobacter* attribution studies performed in Scotland (24) (47), Switzerland (48), New
216 Zealand (49) and Germany (17). This has been informative for understanding the source of human
217 infection but, because of the strong segregation of genetic variation by host (18), it remains possible
218 that collections from multiple countries could be combined to create international isolate
219 collections. This would consolidate research effort and provide the large genome collections
220 necessary for probabilistic attribution models and potential to identify universal host-segregating
221 markers.

222

223 Here we analyzed *C. coli* isolates from Europe and the USA using the conventional MLST method
224 established by Dingle *et al*., in 2001 (27) and specific host-segregating SNPs. A single clonal complex
225 (*CC-828*) dominated among the isolates independently of source and geographical location,
226 representing 780 isolates over 900. The predominance of *CC-828* isolates among *C. coli* (66% - 81%
227 of all isolates (17) (24) (36)) with the *ST-1150* complex accounting for most of the remaining isolates
228 (26), confounds efforts to identify host-association at the clonal complex level – that is possible for
229 *C. jejuni* (18). However, within *CC-828* there was evidence for sequence types that were more
230 commonly isolated from particular hosts. For example, *ST-829*, *ST-832*, *ST-825* and *ST-860*
231 predominated among chicken isolates, *ST-827* was more common in pigs and *ST-1068* was nearly
232 exclusive to cattle, consistent with previous studies (36) (50). Similar low diversity in cattle *C. coli*
233 isolates has previously been described among ruminant isolates from Scotland (47). A weaker host-
234 association signal, based upon MLST alleles, compared to *C. jejuni* has made it difficult to
235 distinctively partition *C. coli* by source (49). However, genotype segregation in *C. coli* provided initial
236 evidence that the genomes of these isolates would contain host-segregating genetic signatures.
237
238 Estimating the discriminating power of genetic markers can be performed by determining the
239 probability that a given genetic element - such as a single mutation - will be found among isolates
240 from a given host (self-attribution). As in previous studies (32) (33), we used STRUCTURE software
241 and self-attribution to determine the predictive power of putative host-segregating markers.
242 Moreover, a recent review (35) mentioned that MLST genes were used for self-attribution tests in 6
243 studies for both *C. coli* and *C. jejuni* (11) (24) (32) (33) (48) (51). However, correct attribution rates
244 for *C. coli* showed inconsistent results for chickens (63-95%), cattle (26-89%) and pigs (70-94%),
245 suggesting that a SNP-based approach may be advantageous for source attribution of *C. coli*. In fact,

246 we showed here that SNPs as host-segregating markers provided more accurate results for chickens,
247 cattle and pigs with 92% (s.d. ± 5.9%), 77% (s.d. ± 8.7%) and 95.3% (s.d. ± 4.4%) correct attribution
248 rates, respectively. While the difficulty in precise self-attribution using MLST genes is undoubtedly
249 linked to reduced resolution, as *CC-828* isolates dominate among *C. coli* populations (23), the
250 transmission of *C. coli* between different host species would also reduce the discriminatory power of
251 source-specific markers potentially leading to incorrect source attribution (22). Adjusting for single
252 mutation determination thus provided promising candidates for accurate source attribution of
253 human *C. coli* isolates. Of 669,019 SNPs from the alignment of 450 genomes against 3 references,
254 259 SNPs in genes associated with cell membrane (transporters, binding proteins), chemotaxis (*FliK*,
255 *FliD*, TLP-like protein), DNA activities (*RecA*, *UvrC*) or energy (*CetC*) functions were chosen for
256 attributing 147 clinical *C. coli* isolates to source.
257
258 It is known that poultry are a major reservoir for human *C. jejuni* infection (8), with a ratio of 9:1 for
259 *C. jejuni* and *C. coli*, respectively (36). Previous studies focussing on the source of *C. coli* infection
260 have come to contrasting conclusions. In France, Sweden, the UK and the USA, the high prevalence
261 of *C. coli* in pigs led to the assumption of the role of this reservoir in human infection (5) (20) (21), up
262 to a ratio of 9:1 in favor of *C. coli* (36). However, in New Zealand, where human *C. coli* infection is
263 also common, there is a low prevalence in pigs (49). Estimates of the relative contribution of
264 different host sources to human infection varies among studies (11) (17) (24) (47) (48) (49) (52) with
265 attribution to poultry (38-86%), ruminants (0-55%) and pigs (1-32%) all being implicated. With the
266 exception of two studies, including rural populations in Scotland and New Zealand, that largely
267 attribute human *C. coli* infections to sheep (47) and ruminants (49), source attribution studies
268 typically assign a principal role for poultry in human infection.

269

270 It is likely that there are differences in the major reservoirs of *C. coli* infection in different countries
271 but quantifying this requires accurate estimation. Estimates based upon MLST loci provided source
272 probabilities with some uncertainty. Specifically, although approximately 40% of the 147 French
273 clinical isolates sampled in this study were clearly attributed (>90% probability, Figure 5), the
274 remaining isolates showed variable scores with many attributed with <60% probability. Overall,
275 MLST-allele-based analyses did assign chicken as a major reservoir for *C. coli* with 89 isolates (61%)
276 attributed with a score equal or greater than 70%. However, this proportion was greatly increased
277 with more accurate attribution scores when using host-segregating SNPs in the attribution model.
278 Specifically, chicken was predicted to be the source of *C. coli* infection for 138 isolates, constituting
279 94% of the clinical samples. In comparison, two recent studies showed that sources of infection of *C.*
280 *jejuni* are more evenly shared between chicken and cattle population in France, with approximately
281 50% for chicken and 40% for cattle, respectively (33) (34). To draw source attribution comparisons
282 between NA and France, additional analyses have been performed using 265 clinical isolates
283 exclusively from the USA (Suppl Figure S2 and Suppl Table S2). The chicken source was again
284 estimated as the main source of *C. coli* contamination in the USA as well as in France, but in a lower
285 proportion (67.9% against 94%) followed by cattle (11.7%) and pig (20.4%). It would be interesting,
286 in a complementary study, to compare the eating habits between these two countries.

In conclusion, the added resolution provided by genome-wide host-segregating markers not only improves source attribution for *C. coli* but also provides important information about the major infection reservoirs that has been missed in some previous studies (21). By combining whole genome analysis with national surveillance programs and source attribution modelling it was possible to identify the chicken reservoir as a major source of *C. coli* infection in France and abroad. These findings will support ongoing surveillance and the development of targeted interventions aimed at reducing the burden of human campylobacteriosis.

Applied and Environmental Microbiology

294 **MATERIAL AND METHODS**

295 ***Campylobacter coli* isolate datasets**

296 A total of 450 *C. coli* isolates genomes from two major regions where Campylobacters are a leading
297 cause of foodborne infections, North America and Europe, were selected for the determination of
298 host-segregating markers (Dataset S1). To reduce the detection of regional-specific markers, these
299 genomes were randomly selected from multiple countries within these two regions. That included
300 even numbers (n=150) of chicken, cattle and pig *C. coli* genomes to avoid bias in the identification of
301 host-specific markers. This first dataset was comprised of 151 isolates from PubMLST databases (31)
302 and 299 from the USA National Antimicrobial Resistance Monitoring System (NARMS) project (53).
303 PubMLST genomes comprised 34% of that first dataset and included 47%, 7% and 47% of all chicken,
304 cattle and pig marker-determination isolates, respectively. NARMS genomes comprised 66% of the
305 dataset, and included 53%, 93% and 53% of all chicken, cattle and pig marker-determination
306 isolates. These datasets were entirely composed of European and North American genomes.
307 European isolates represented 29% of the dataset, including 41%, 1% and 45% of all chicken, cattle
308 and pig isolates respectively, while North American isolates comprised 71% of the dataset including
309 59%, 99% and 55% of all chicken, cattle and pigs isolates. North American isolates were mostly
310 selected from the USA (n=315). Remaining isolates (n=4) were selected from Canada. A total of 424
311 isolates (94%) were obtained from 2005 to 2019.
312
313 A second dataset (validation dataset) comprised 300 supplementary *C. coli* isolates of known source
314 reservoirs was used in order to test the discriminatory strength the host-segregating SNPs previously
315 obtained (Dataset S2). This dataset comprised North American *C. coli* isolates from the NARMS
316 project; 100 for each source. Finally, 150 French clinical isolates comprised a last set of genomes

(clinical dataset) and were used to attribute the putative source reservoir of clinical isolates (Dataset S3). This comprised 150 clinical isolates from French laboratories and hospitals surveillance network sampled from stools between 2015 to 2017. Clinical isolates were chosen to represent patients from diverse geographic regions in France, with a sex ratio of 1.03 and a mean age of 39.4 s.d. ±2.8 years old.

Clinical isolate genomes had an average genome length of 1.7 Mbp (s.d. ± 69.7 Kbp) and an average number of contigs of 43. *C. coli* marker-determination isolates were on average 1.76 Mbp (s.d. ± 81.2 Kbp) in length and comprised 83 contigs; and *C. coli* validation isolates were on average 1.78 Mbp (s.d. ± 74.7 Kbp) in length over 78 contigs (Suppl Figure S1). This is consistent with other published *C. coli* genomes, estimated to ~1.7 Mbp in length (54). Furthermore, no significant difference in *C. coli* genome sizes from different hosts has been observed: *C. coli* isolated from chickens were on average 1.78 Mbp in length (s.d. ± 106 Kbp), 1.77 Mbp (s.d. ± 61.6 Kbp) for cattle isolates and 1.77 Mbp (s.d. ± 61 Kbp) for pig isolates.

DNA extraction, genome sequencing and assembly

DNA from clinical isolates was extracted using the MagNA Pure 6 DNA and Viral NA SV Kit and DNA purification was performed from bacterial lysis on a MagNA Pure 96 System (Roche Applied Science, Manheim, Germany). Quantification and purity checks (260/280 and 260/230 ratios) were determined by spectrophotometry (NanoDrop Technologies, Wilmington, DE, USA) before sequencing. Paired-end next-generation sequencing was performed on DNA samples using Illumina HiSeq 4000 technology (Integragen, Evry, France). Additionally, FastQC v0.11.8 (55) was used to run data quality tests. Genomic data was cleaned and genomes were assembled using Sickle v1.33 (56)

and SPAdes v3.10.1 (57), respectively. Genomes were then filtered in order to remove poor quality contigs: sequences with a length smaller than 160 nucleotides and a k-mer coverage less than 20x were removed. One isolate (2015_0475) showed an abnormal genome size of 2.5 Mbp after filtration and was excluded from subsequent analyses.

Characterization of genomic variation

In silico, MLST was performed for a comparative analysis with host-segregating SNPs. Profiles were obtained for all 900 isolates using 7 housekeeping genes (*aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt* and *uncA*) determined for *Campylobacter* species (27). Sequence types (STs) and clonal complexes ("CC", groups of isolates with a sequence type that share four or more loci (27)) were defined using the sequence tag tool of PubMLST (58). Using this method, two clinical isolates (2016_1990 and 2017_2288) and one validation isolate (FSIS11705596) were miss-identified as *C. coli* and were actually *C. jejuni* and removed from the dataset. The updated validation and clinical datasets were then comprised of 299 and 147 isolates, respectively. A phylogenetic tree was constructed according to all sequence types using GrapeTree (59). A second tree was built based on every host-segregating markers determined in this study, in order to make a direct comparison with the MLST tree. A multi-fasta file containing sequences from concatenated SNPs of all isolates (n=896) was created. Sequences were aligned using Muscle v3.8.1551 (60) and a Newick format tree from maximum-likelihood method was generated using Fasttree v2.1.11 (61). Microreact online platform was used to visualize the tree (62).

To identify candidate SNPs, genome-wide variant calling was primarily performed by aligning all isolates from the marker-determination dataset (n=450) to *C. coli* reference genomes. Three

363 references from each source were chosen in order to target source-specific genomic regions and
364 capture all potential markers: OR12 strain isolated from a chicken (NZ_CP019977.1) (63), HC2-48
365 strain isolated from a cow (NZ_CP013034.1) (64) and ZV1224 strain isolated from a pig
366 (NZ_CP017875.1) (65). The bwa v0.7.17 (66) tool developed for mapping sequences against given
367 genomes was used here to align each isolate to OR12, HC2-48 and ZV1224 references. Alignment
368 files were sorted using samtools v1.9 (67). Genotypes were determined with bcftools v1.9 "mpileup"
369 variant calling tool (67), and 3 variant calling files (vcf) were generated (one for each reference). A
370 script was written in Python (see data availability) to filter all SNP variations found in more than 2
371 out of 3 isolates. Since a source represents 33% of the total dataset (150 isolates over 450), a
372 proportion greater than 66% means that a same SNP variation is likely to be found in each of the 3
373 selected sources. Therefore, this step enabled the removal of weakly discriminating polymorphisms
374 and reduced the computational time of subsequent analyses.
375
376 **Identification of host-segregating marker**
377 In order to identify host-segregating markers, source attribution tests of marker-determination
378 isolates (of known sources) were performed using all previously selected SNPs individually to
379 identify host-segregating markers. A matrix was constructed of all genotypes in the 450 marker-
380 determination isolate dataset (nucleotides were translated into numbers: "1" for "A", "2" for "T",
381 etc.). Source attribution tests were performed in triplicate for each SNP using STRUCTURE (68), with
382 the no admixture model, 3 putative populations (K = 3), 10,000 iterations, and a burn-in period of
383 10,000 iterations. For each STRUCTURE test, 60 different random isolates (20 from each population)
384 were set to "unknown source" (POPFLAG = 0) in order to estimate the probability of correct self-
385 attribution, and then to evaluate the SNP host-segregating strength. Each SNP with 70% or greater

of total correct self-attributions for at least one source was selected; a minimum of 66% (here rounded up to 70%) of source attribution rate indicates that a variant is discriminating between at least 1 out of 3 sources. Additionally, genomic sequences containing the selected SNPs were extracted from the corresponding reference (OR12, HC2-48 or ZV1224) and annotated using blast-x online tool (69).

Validation of the discriminatory power of host-segregating markers

To validate the capability of the selected SNPs to discriminate isolates from different populations, STRUCTURE tests were run again using the marker-determination dataset and different sets of markers: SNPs contained in the same CDS, all SNPs determined from OR12, HC2-48 and ZV1224 alignments and all SNPs from all alignments. One hundred tests were then performed using each set of SNPs and 60 random isolates per test for self-attribution (POPFLAG = 0) ("no admixture model", K = 3, 10,000 iterations and a burn-in period of 10,000 iterations). Additionally, source attribution of 299 validation isolates of known source reservoirs, which were not used for SNP determination, was performed. Specifically, each SNP was obtained using samtools mpileup option. STRUCTURE was run 10 times using marker-determination isolates as training dataset (n=450) and validation dataset as unknown source isolates (POPFLAG = 0). STRUCTURE model parameters remained unchanged. Each validation isolate was attributed to its source based on the average of attribution rate of all 10 tests. An isolate was considered correctly source re-attributed with a STRUCTURE score greater than 70%. In each case the same method was performed simultaneously with MLST alleles to compare the discriminating strength of both type of markers (SNP or allele).

Source attribution of clinical isolates

Similar to validation analysis, source attribution of *C. coli* clinical isolates was performed using determined host-segregating markers thus in order to identify the main source of infection in France. For each SNP (n=259), every genotype was extracted from all clinical isolates using samtools mpileup option. STRUCTURE was run 10 times using marker-determination isolates as training dataset (n=450) and clinical dataset (n=147) as unknown source isolates (POPFLAG = 0) (K = 3, 10,000 iterations and a burn-in period of 10,000 iterations). Each clinical isolate was attributed to a source based on the average of attribution rate of all 10 tests. Source attribution of clinical isolates was performed simultaneously with MLST alleles to compare proportions of each source between both type of markers (SNP or allele).

DATA AVAILABILITY

All 900 *C. coli* genomes are available using IDs listed in Dataset S1, S2 and S3: BioSample and PubMLST IDs for NCBI and PubMLST databases respectively.

Personnal VCF filter Python script available on GitHub: QuentinJehanne. (2020, April 8). QuentinJehanne/ccoli_2020: v1 of a personal VCF filter (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.3744758.

ACKNOWLEDGMENT

The authors want to thank all the laboratories that sent *Campylobacter* isolates to our reference centre. This study was financed by internal funding of the French National Reference Center for Campylobacters and Helicobacters (Bordeaux, France) (www.cnrch.fr). The material is original research and has not been previously published nor submitted for publication elsewhere. The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

Conceptualization, Quentin Jehanne and Philippe Lehours; Funding acquisition, Philippe Lehours; Investigation, Quentin Jehanne, Ben Pascoe and Samuel K Sheppard; Project administration, Quentin Jehanne, Samuel K Sheppard and Philippe Lehours; Resources, Lucie Bénéjat, Astrid Ducournau, Alice Buissonnière, Evangelos Mourkas, Ben Pascoe and Samuel K Sheppard; Supervision, Philippe Lehours and Samuel K Sheppard; Validation, Quentin Jehanne, Samuel K Sheppard and Philippe Lehours; Visualization, Quentin Jehanne, Ben Pascoe, Samuel K Sheppard and Philippe Lehours; Writing – original draft, Quentin Jehanne and Philippe Lehours; Writing – review & editing, Quentin

442 Jehanne, Samuel K Sheppard, Ben Pascoe, Francis Mégraud, Emilie Bessède and Philippe Lehours. All

443 authors have read and agreed to the published version of the manuscript.

444 **REFERENCES**

445 (1) Blaser MJ. 1997. Epidemiologic and clinical features of *Campylobacter jejuni* infections. J
446 Infect Dis 176 Suppl 2:S103-105.

447 (2) Scallan E, Griffin PM, Angulo FJ, Tauxe RV, Hoekstra RM. 2011. Foodborne illness acquired in
448 the United States--unspecified agents. Emerging Infect Dis 17:16–22.

449 (3) The European Union summary report on trends and sources of zoonoses, zoonotic agents
450 and food-borne outbreaks in 2017 - - 2018 - EFSA Journal - Wiley Online Library.

451 (4) Van Cauteren D, Le Strat Y, Sommen C, Bruyand M, Tourdjman M, Da Silva NJ, Couturier E,
452 Fournet N, de Valk H, Desenclos J-C. 2017. Estimated Annual Numbers of Foodborne
453 Pathogen-Associated Illnesses, Hospitalizations, and Deaths, France, 2008-2013. Emerging
454 Infect Dis 23:1486–1492.

455 (5) Horrocks SM, Anderson RC, Nisbet DJ, Ricke SC. 2009. Incidence and ecology of
456 *Campylobacter jejuni* and *coli* in animals. Anaerobe 15:18–25.

457 (6) French National Reference Center for Campylobacters & Helicobacters (Bordeaux Hospital
458 University Center). 2019. 2018 Campylobacters surveillance report.

459 (7) Fitzgerald C. 2015. Campylobacter. Clin Lab Med 35:289–298.

460 (8) Kapperud G, Rosef O. 1983. Avian wildlife reservoir of *Campylobacter fetus subsp*. *jejuni*,
461 *Yersinia spp.*, and *Salmonella spp*. in Norway. Appl Environ Microbiol 45:375–380.

462 (9) Sheppard SK, Dallas JF, MacRae M, McCarthy ND, Sproston EL, Gormley FJ, Strachan NJC,
463 Ogden ID, Maiden MCJ, Forbes KJ. 2009. Campylobacter genotypes from food animals,
464 environmental sources and clinical disease in Scotland 2005/6. Int J Food Microbiol 134:96–
465 103.

466 (10) Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, Fox A,
467 Fearnhead P, Hart CA, Diggle PJ. 2008. Tracing the source of campylobacteriosis. PLoS Genet
468 4:e1000203.
469 (11) Sheppard SK, Dallas JF, Strachan NJC, MacRae M, McCarthy ND, Wilson DJ, Gormley
470 FJ, Falush D, Ogden ID, Maiden MCJ, Forbes KJ. 2009. Campylobacter genotyping to
471 determine the source of human infection. Clin Infect Dis 48:1072–1078.
472 (12) Nachamkin I, Allos BM, Ho T. 1998. *Campylobacter* species and Guillain-Barré
473 syndrome. Clin Microbiol Rev 11:555–567.
474 (13) Grover M. 2014. Role of gut pathogens in development of irritable bowel syndrome.
475 Indian J Med Res 139:11–18.
476 (14) Yang Y, Feye KM, Shi Z, Pavlidis HO, Kogut M, J. Ashworth A, Ricke SC. 2019. A
477 Historical Review on Antibiotic Resistance of Foodborne Campylobacter. Front Microbiol
478 10:1509.
479 (15) Salazar-Lindo E, Sack RB, Chea-Woo E, Kay BA, Piscoya ZA, Leon-Barua R, Yi A. 1986.
480 Early treatment with erythromycin of *Campylobacter jejuni*-associated dysentery in children.
481 J Pediatr 109:355–360.
482 (16) Mourkas E, Florez-Cuadrado D, Pascoe B, Calland JK, Bayliss SC, Mageiros L, Méric G,
483 Hitchings MD, Quesada A, Porrero C, Ugarte-Ruiz M, Gutiérrez-Fernández J, Domínguez L,
484 Sheppard SK. 2019. Gene pool transmission of multidrug resistance among Campylobacter
485 from livestock, sewage and human disease. Environmental Microbiology 21:4597–4613.
486 (17) Rosner BM, Schielke A, Didelot X, Kops F, Breidenbach J, Willrich N, Gölz G, Alter T,
487 Stingl K, Josenhans C, Suerbaum S, Stark K. 2017. A combined case-control and molecular

source attribution study of human Campylobacter infections in Germany, 2011–2014. Sci Rep 7:1–12.

(18) Sheppard SK, Colles F, Richardson J, Cody AJ, Elson R, Lawson A, Brick G, Meldrum R, Little CL, Owen RJ, Maiden MCJ, McCarthy ND. 2010. Host association of Campylobacter genotypes transcends geographic variation. Appl Environ Microbiol 76:5269–5277.

(19) Pascoe B, Méric G, Yahara K, Wimalarathna H, Murray S, Hitchings MD, Sproston EL, Carrillo CD, Taboada EN, Cooper KK, Huynh S, Cody AJ, Jolley KA, Maiden MCJ, McCarthy ND, Didelot X, Parker CT, Sheppard SK. 2017. Local genes for local bacteria: Evidence of allopatry in the genomes of transatlantic Campylobacter populations. Mol Ecol 26:4497–4508.

(20) Ogden ID, Dallas JF, MacRae M, Rotariu O, Reay KW, Leitch M, Thomson AP, Sheppard SK, Maiden M, Forbes KJ, Strachan NJC. 2009. Campylobacter Excreted into the Environment by Animal Sources: Prevalence, Concentration Shed, and Host Association. Foodborne Pathog Dis 6:1161–1170.

(21) Kempf I, Kerouanton A, Bougeard S, Nagard B, Rose V, Mourand G, Osterberg J, Denis M, Bengtsson BO. 2017. *Campylobacter coli* in Organic and Conventional Pig Production in France and Sweden: Prevalence and Antimicrobial Resistance. Front Microbiol 8:955.

(22) Dearlove BL, Cody AJ, Pascoe B, Méric G, Wilson DJ, Sheppard SK. 2016. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. ISME J 10:721–729.

(23) Thakur S, Morrow WEM, Funk JA, Bahnson PB, Gebreyes WA. 2006. Molecular Epidemiologic Investigation of *Campylobacter coli* in Swine Production Systems, Using Multilocus Sequence Typing. Appl Environ Microbiol 72:5666–5669.

510 (24) Sheppard SK, Dallas JF, Wilson DJ, Strachan NJC, McCarthy ND, Jolley KA, Colles FM,
511 Rotariu O, Ogden ID, Forbes KJ, Maiden MCJ. 2010. Evolution of an agriculture-associated
512 disease causing *Campylobacter coli* clade: evidence from national surveillance data in
513 Scotland. PLoS ONE 5:e15708.
514 (25) Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. 2008. Convergence of
515 Campylobacter species: implications for bacterial evolution. Science 320:237–239.
516 (26) Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A,
517 Colles FM, Strachan NJC, Ogden ID, Forbes K, French NP, Carter P, Miller WG, McCarthy ND,
518 Owen R, Litrup E, Egholm M, Affourtit JP, Bentley SD, Parkhill J, Maiden MCJ, Falush D. 2013.
519 Progressive genome-wide introgression in agricultural *Campylobacter coli*. Mol Ecol 22:1051–
520 1064.
521 (27) Dingle KE, Colles FM, Wareing DRA, Ure R, Fox AJ, Bolton FE, Bootsma HJ, Willems RJL,
522 Urwin R, Maiden MCJ. 2001. Multilocus Sequence Typing System for *Campylobacter jejuni*. J
523 Clin Microbiol 39:14–23.
524 (28) Clark CG, Bryden L, Cuff WR, Johnson PL, Jamieson F, Ciebin B, Wang G. 2005. Use of
525 the Oxford Multilocus Sequence Typing Protocol and Sequencing of the Flagellin Short
526 Variable Region To Characterize Isolates from a Large Outbreak of Waterborne
527 *Campylobacter* sp. Strains in Walkerton, Ontario, Canada. J Clin Microbiol 43:2080–2091.
528 (29) Sheppard SK, Cheng L, Méric G, Haan CPA de, Llarena A-K, Marttinen P, Vidal A, Ridley
529 A, Clifton-Hadley F, Connor TR, Strachan NJC, Forbes K, Colles FM, Jolley KA, Bentley SD,
530 Maiden MCJ, Hänninen M-L, Parkhill J, Hanage WP, Corander J. 2014. Cryptic ecology among
531 host generalist *Campylobacter jejuni* in domestic animals. Molecular Ecology 23:2442–2451.

532 (30) Gripp E, Hlahla D, Didelot X, Kops F, Maurischat S, Tedin K, Alter T, Ellerbroek L,
533 Schreiber K, Schomburg D, Janssen T, Bartholomäus P, Hofreuter D, Woltemate S, Uhr M,
534 Brenneke B, Grüning P, Gerlach G, Wieler L, Suerbaum S, Josenhans C. 2011. Closely related
535 *Campylobacter jejuni* strains from different sources reveal a generalist rather than a
536 specialist lifestyle. BMC Genomics 12:584.
537 (31) Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics:
538 BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res
539 3:124.
540 (32) Thépault A, Méric G, Rivoal K, Pascoe B, Mageiros L, Touzain F, Rose V, Béven V,
541 Chemaly M, Sheppard SK. 2017. Genome-Wide Identification of Host-Segregating
542 Epidemiological Markers for Source Attribution in *Campylobacter jejuni*. Appl Environ
543 Microbiol 83:e03085-16.
544 (33) Thépault A, Rose V, Quesne S, Poezevara T, Béven V, Hirchaud E, Touzain F, Lucas P,
545 Méric G, Mageiros L, Sheppard SK, Chemaly M, Rivoal K. 2018. Ruminant and chicken:
546 important sources of campylobacteriosis in France despite a variation of source attribution in
547 2009 and 2015. Sci Rep 8:1–10.
548 (34) Berthenet E, Thépault A, Chemaly M, Rivoal K, Ducournau A, Buissonnière A, Bénéjat
549 L, Bessède E, Mégraud F, Sheppard SK, Lehours P. 2019. Source attribution of *Campylobacter*
550 *jejuni* shows variable importance of chicken and ruminants reservoirs in non-invasive and
551 invasive French clinical isolates. Sci Rep 9:8098.
552 (35) Cody AJ, Maiden MC, Strachan NJ, McCarthy ND. 2019. A systematic review of source
553 attribution of human campylobacteriosis using multilocus sequence typing. Euro Surveill 24.

554 (36) Miller WG, Englen MD, Kathariou S, Wesley IV, Wang G, Pittenger-Alley L, Siletz RM,
555 Muraoka W, Fedorka-Cray PJ, Mandrell RE. 2006. Identification of host-associated alleles by
556 multilocus sequence typing of *Campylobacter coli* strains from food animals. Microbiology
557 (Reading, Engl) 152:245–255.
558 (37) Kamal N, Dorrell N, Jagannathan A, Turner SM, Constantinidou C, Studholme DJ,
559 Marsden G, Hinds J, Laing KG, Wren BW, Penn CW. 2007. Deletion of a previously
560 uncharacterized flagellar-hook-length control gene fliK modulates the sigma54-dependent
561 regulon in *Campylobacter jejuni*. Microbiology (Reading, Engl) 153:3099–3111.
562 (38) Yeh H-Y, Hiett KL, Line JE, Seal BS. 2014. Characterization and antigenicity of
563 recombinant *Campylobacter jejuni* flagellar capping protein FliD. J Med Microbiol 63:602–
564 609.
565 (39) Clark C, Berry C, Demczuk W. 2019. Diversity of transducer-like proteins (Tlps) in
566 Campylobacter. PLoS ONE 14:e0214228.
567 (40) Reuter M, HM van Vliet A. Signal Balancing by the CetABC and CetZ Chemoreceptors
568 Controls Energy Taxis in *Campylobacter jejuni*. PLoS One 8:e54390.
569 (41) Runti G, Ruiz M del CL, Stoilova T, Hussain R, Jennions M, Choudhury HG, Benincasa
570 M, Gennaro R, Beis K, Scocchi M. 2013. Functional Characterization of SbmA, a Bacterial
571 Inner Membrane Transporter Required for Importing the Antimicrobial Peptide Bac7(1-35).
572 Journal of Bacteriology 195:5343–5351.
573 (42) Elvers KT, Wu G, Gilberthorpe NJ, Poole RK, Park SF. 2004. Role of an inducible single-
574 domain hemoglobin in mediating resistance to nitric oxide and nitrosative stress in
575 *Campylobacter jejuni* and *Campylobacter coli*. J Bacteriol 186:5332–5341.

576 (43) Bertrand-Burggraf E, Selby CP, Hearst JE, Sancar A. 1991. Identification of the
577 different intermediates in the interaction of (A)BC excinuclease with its substrates by DNase I
578 footprinting on two uniquely modified oligonucleotides. J Mol Biol 219:27–36.
579 (44) Sheppard SK, Guttman DS, Fitzgerald JR. 2018. Population genomics of bacterial host
580 adaptation. 9. Nat Rev Genet 19:549–565.
581 (45) Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden
582 MCJ, Parkhill J, Falush D. 2013. Genome-wide association study identifies vitamin B5
583 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci USA 110:11923–
584 11927.
585 (46) Yahara K, Méric G, Taylor AJ, Vries SPW de, Murray S, Pascoe B, Mageiros L, Torralbo
586 A, Vidal A, Ridley A, Komukai S, Wimalarathna H, Cody AJ, Colles FM, McCarthy N, Harris D,
587 Bray JE, Jolley KA, Maiden MCJ, Bentley SD, Parkhill J, Bayliss CD, Grant A, Maskell D, Didelot
588 X, Kelly DJ, Sheppard SK. 2017. Genome-wide association of functional traits linked with
589 *Campylobacter jejuni* survival from farm to fork. Environmental Microbiology 19:361–380.
590 (47) Roux F, Sproston E, Rotariu O, Macrae M, Sheppard SK, Bessell P, Smith-Palmer A,
591 Cowden J, Maiden MCJ, Forbes KJ, Strachan NJC. 2013. Elucidating the aetiology of human
592 *Campylobacter coli* infections. PLoS ONE 8:e64504.
593 (48) Kittl S, Heckel G, Korczak BM, Kuhnert P. 2013. Source Attribution of Human
594 Campylobacter Isolates by MLST and Fla-Typing and Association of Genotypes with
595 Quinolone Resistance. PLoS One 8:e81796.
596 (49) Nohra A, Grinberg A, Midwinter AC, Marshall JC, Collins-Emerson JM, French NP.
597 2016. Molecular Epidemiology of *Campylobacter coli* Strains Isolated from Different Sources
598 in New Zealand between 2005 and 2014. Appl Environ Microbiol 82:4363–4370.

599 (50) Rotariu O, Dallas JF, Ogden ID, MacRae M, Sheppard SK, Maiden MCJ, Gormley FJ,
600 Forbes KJ, Strachan NJC. 2009. Spatiotemporal Homogeneity of Campylobacter Subtypes
601 from Cattle and Sheep across Northeastern and Southwestern Scotland. Appl Environ
602 Microbiol 75:6275–6281.
603 (51) Smid JH, Gras LM, Boer AG de, French NP, Havelaar AH, Wagenaar JA, Pelt W van.
604 2013. Practicalities of Using Non-Local or Non-Recent Multilocus Sequence Typing Data for
605 Source Attribution in Space and Time of Human Campylobacteriosis. PLOS ONE 8:e55029.
606 (52) Mossong J, Mughini-Gras L, Penny C, Devaux A, Olinger C, Losch S, Cauchie H-M, van
607 Pelt W, Ragimbeau C. 2016. Human Campylobacteriosis in Luxembourg, 2010–2013: A Case-
608 Control Study Combined with Multilocus Sequence Typing for Source Attribution and Risk
609 Factor Analysis. Sci Rep 6:20939.
610 (53) Karp BE, Tate H, Plumblee JR, Dessai U, Whichard JM, Thacker EL, Hale KR, Wilson W,
611 Friedman CR, Griffin PM, McDermott PF. 2017. National Antimicrobial Resistance Monitoring
612 System: Two Decades of Advancing Public Health Through Integrated Surveillance of
613 Antimicrobial Resistance. Foodborne Pathog Dis 14:545–557.
614 (54) Pearson BM, Rokney A, Crossman LC, Miller WG, Wain J, van Vliet AHM. 2013.
615 Complete Genome Sequence of the *Campylobacter coli* Clinical Isolate 15-537360. Genome
616 Announc 1:e01056-13.
617 (55) Wingett SW, Andrews S. 2018. FastQ Screen: A tool for multi-genome mapping and
618 quality control. F1000Res 7:1338.
619 (56) Joshi N, Fass J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool
620 for FastQ files (Version 1.33)[Software].

621 (57) Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
622 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev
623 MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to
624 Single-Cell Sequencing. J Comput Biol 19:455–477.
625 (58) Jolley KA, Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation
626 at the population level. BMC Bioinformatics 11:595.
627 (59) Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA,
628 Achtman M. 2018. GrapeTree: visualization of core genomic relationships among 100,000
629 bacterial pathogens. Genome Res 28:1395–1404.
630 (60) Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
631 throughput. Nucleic Acids Res 32:1792–1797.
632 (61) Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing Large Minimum Evolution
633 Trees with Profiles instead of a Distance Matrix. Mol Biol Evol 26:1641–1650.
634 (62) Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden
635 MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and
636 sharing data for genomic epidemiology and phylogeography. Microb Genom 2:e000093.
637 (63) O'Kane PM, Connerton IF. 2017. Characterisation of Aerotolerant Forms of a Robust
638 Chicken Colonizing *Campylobacter coli*. Front Microbiol 8:513.
639 (64) Marasini D, Fakhr MK. 2016. Complete Genome Sequences of the Plasmid-Bearing
640 *Campylobacter coli* Strains HC2-48, CF2-75, and CO2-160 Isolated from Retail Beef Liver.
641 Genome Announc 4(5):e1004-16.

(65) Marasini D, Fakhr MK. 2017. Complete Genome Sequences of Plasmid-Bearing Multidrug-Resistant *Campylobacter jejuni* and *Campylobacter coli* Strains with Type VI Secretion Systems, Isolated from Retail Turkey and Pork. Genome Announc 5(47):e01360-17.

(66) Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv.

(67) Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

(68) Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

(69) Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

655 **Tables**

656

657 **Table 1. Variant calling comparison between 3 references of *C. coli***

658

659 **Table 2. Rates of correct self-attributions of marker-determination isolates using 5 different set of**

660 **markers**

661

662 **Table 3. Source attribution scores of clinical isolates**

663 **Figure Legends**

664

665 **Figure 1: Phylogenic tree based on MLST analysis**

666 The minimum spanning tree was generated using GrapeTree from the sequence types of all 896 *C.*
667 *coli* isolates, based on 7 MLST genes (*aspA*, *glnA*, *gltA*, *glyA*, *pgm*, *tkt* and *uncA*) extracted using
668 PubMLST platform. Orange color represents isolates isolated from chickens, green color from cattle,
669 magenta color from pigs and red color are for clinical isolates. Circle sizes are proportional to the
670 number of isolates and the scale bar represents a genetic distance of 1.

671

672 **Figure 2: Phylogenic tree built from concatenated selected SNPs**

673 Tree designed using maximum-likelihood phylogeny between 896 isolate sequences built from the
674 concatenation of all genotypes of the selected SNPs (n=259). Orange nodes are the chicken
675 population isolates, green nodes for cattle isolates, pink nodes for pig isolates and red nodes for
676 clinical isolates. Orange circle shows an estimation of the chicken cluster, green circle for the cattle
677 cluster and pink circle for the pig cluster and the scale bar represents a genetic distance of 0.24.
678 Clinical isolates are mostly located within the chicken cluster, which is consistent with the
679 probabilistic attribution model.

680

681 **Figure 3: Host-segregating rate of all variants obtained from the alignment of 450 marker-**
682 **determination isolates against 3 references**

683 Source attribution rates (y axes) were obtained testing 26,131, 24,395 and 20,827 SNPs from OR12
684 (**a**), HC2-48 (**b**) and ZV1224 (**c**) references, respectively, and are shown here according to their
685 genome position (left, x axis) and variant proportions (right, x axis). STRUCTURE software was run 3

686 times for each SNP (average attribution rates are shown here), using 390 *C. coli* randomly selected
687 isolates as training dataset and 60 randomly selected isolates as test dataset. Orange color
688 represents attribution rates and number of SNPs for chicken source, green color for cattle source
689 and magenta color for pig source. A total of 259 SNPs showed attribution rates greater than 70%
690 (red line) for one or more sources and were carried forward for further analyses: 43, 183 and 33
691 SNPs from chicken, cattle and pig references, respectively. Scores fluctuated between 30% and 40%
692 and highest attribution rates for each host reservoir were found in the corresponding source
693 reference. However, OR12 reference showed two distinct regions of the genome: one part
694 containing variants discriminating the chicken source and another part the pig source. Two low
695 variable regions (blanks), where no SNP from the variant calling step were selected, are also visible.
696
697 **Figure 4: Correct re-attribution proportions of 299 validation isolates using determined SNPs and**
698 **MLST genes**
699 Source attribution strength of selected SNPs (**a**) and MLST genes (**b**) estimated using STRUCTURE
700 software. A total of 299 isolates were tested (from the validation dataset) using marker-
701 determination isolates for training (n=450). Source attributions were performed 10 times using all
702 selected SNPs (n=259) and MLST genes (n=7). Gray bars represent rate of correct source attribution
703 for chicken population isolates, black bars for cattle isolates and white bars for pig isolates. An
704 isolate was considered correctly source re-attributed with a STRUCTURE score greater than 70%.
705
706 **Figure 5: Population proportions of clinical isolates from source attribution**
707 Source attribution of clinical dataset using selected SNPs (**a**, n=259) and MLST genes (**b**, n=7). Clinical
708 isolates (n=147) are represented on x axis and their attribution probabilities on y axis in orange for

chicken source, green for cattle source and pink for pig source. The poultry reservoir was estimated as the main source of *C. coli* contamination in France with 138 isolates (94%) attributed using host-segregating SNPs and 89 isolates (61%) using MLST (isolates selected with source probabilities greater than 70%).

713 **Supplementary Tables and Figures**

714

715 **Dataset S1. Marker-determination dataset isolates**

716

717 **Dataset S2. Validation dataset isolates**

718

719 **Dataset S3. Clinical dataset isolates**

720

721 **Suppl Figure S1. WGS data from all 900 *C. coli* isolates.**

722

723 **Suppl Table S1. List of all determined proteins with their corresponding number of SNPs.**

724

725 **Suppl Figure S2. Population proportions from source attribution of 265 *C. coli* clinical isolates from**

726 **the USA.**

727

728 **Suppl Table S2. List of all *C. coli* isolates from the USA selected for source attribution.**

729

730

Table 1. Variant calling comparison between 3 references of *Campylobacter coli*

Reference	Variant calling raw[1]	Filtration[2]	Selected SNPs[3]
OR12 (chicken)	283,320	26,131	43
HC2-48 (cattle)	202,111	24,395	183
ZV1224 (pig)	242,574	20,827	33

[1]Number of SNPs determined after aligning all isolates from marker-determination (n=450) dataset to 3 different references of *C. coli* : OR12 isolated from chicken, HC2-48 from cattle and ZV1224 from pig.

[2]Number of SNPs after the filtration of genotypes which represent more than two third of all isolates.

[3]Selected SNPs with 70% or greater of total correct self-attributions.

Table 2. Rates of correct self-attributions of marker-determination isolates using 5 different set of markers

	Chicken isolates self-attributions (n = 150)		**Cattle isolates self-attributions (n = 150)**		**Pig isolates self-attributions (n = 150)**	
Set of markers	Rate of correct attribution (%)	Std deviation (%)	Rate of correct attribution (%)	Std deviation (%)	Rate of correct attribution (%)	Std deviation (%)
43 SNPs (OR12)	88.4	± 6.24	63.8	± 9.22	96.2	± 4.09
183 SNPs (HC2-48)	91.1	± 5.74	75.0	± 9.69	*42.5*	± *18.74*
33 SNPs (ZV1224)	75.0	± 13.9	*19.7*	± *10.08*	94.7	± 5.23
259 SNPs (all)	**92.0**	**± 5.86**	**77.0**	**± 8.65**	**95.3**	**± 4.4**
7 genes (MLST)	**73.6**	**± 9.06**	**76.8**	**± 9.42**	**74.4**	**± 9.50**

Discriminating strength of selected SNPs and MLST genes were estimated using marker-determination isolates. From 450 initial isolates, random selections of 390 and 60 isolates were used for training and self-attribution (sources set to "unknown"), respectively. Self-attributions were performed 100 times using selected SNPs from chicken alignment (n=43), from cattle alignment (n=183), from pig alignment (n=33), from all alignments (n=259) and 50 times using MLST genes (n=7). Since multiple tests were performed for each set of markers using 60 randomly selected isolates, standard deviations were calculated.

Table 3. Source attribution scores of clinical isolates

	Attribution to chicken source	Attribution to cattle source	Attribution to pig source	Inconclusive attributions
Set of markers	% of clinical isolates (Average score %)	% of clinical isolates (Average score %)	% of clinical isolates (Average score %)	% of clinical isolates (Average score %)
259 SNPs	93.88 (100.0)	0.0 (0.0)	6.12 (100.0)	0.0 (0.0)
7 MLST genes	60.54 (88.35)	8.84 (86.91)	4.08 (83.22)	26.53 (50.59)

Data for source attribution of clinical isolates dataset using selected SNPs (n=259) and MLST genes (n=7). "% *of clinical isolates*" show the distribution of estimated sources among clinical isolates with "*Average score*" as their average of individual attribution rate. Using determined SNPs, source attribution rates for clinical isolates were constant whereas using MLST genes, source attribution showed variable results.